

# Evolution of genome size in asexual populations

Aditi Gupta<sup>\*1,2</sup>, Thomas LaBar<sup>1,2</sup>, Michael Miyagi<sup>3</sup>, and Christoph Adami<sup>1,2,4</sup>

<sup>1</sup>BEACON Center for the Study of Evolution in Action, Michigan State University, East Lansing, MI 48824

<sup>2</sup>Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI 48824

<sup>3</sup>Department of Integrative Biology, the University of Texas at Austin, Austin, TX

<sup>4</sup>Department of Physics and Astronomy, Michigan State University, East Lansing, MI 48824

## Abstract

Genome sizes have evolved to vary widely, from 250 bases in viroids to 670 billion bases in amoeba. This remarkable variation in genome size is the outcome of complex interactions between various evolutionary factors such as point mutation rate, population size, insertions and deletions, and genome editing mechanisms that may be specific to certain taxonomic lineages. While comparative genomics analyses have uncovered some of the relationships between these diverse evolutionary factors, we still do not understand what drives genome size evolution. Specifically, it is not clear how primordial mutational processes of base substitutions, insertions, and deletions influence genome size evolution in asexual organisms. Here, we use digital evolution to investigate genome size evolution by tracking genome edits and their fitness effects in real time. In agreement with empirical data, we find that mutation rate is inversely correlated with genome size in asexual populations. We show that at low point mutation rate, insertions are significantly more beneficial than deletions, driving genome expansion and acquisition of phenotypic complexity. Conversely, high mutational load experienced at high mutation rates inhibits genome growth, forcing the genomes to compress genetic information. Our analyses suggest that the inverse relationship between mutation rate and genome size is a result of the tradeoff between evolving phenotypic innovation and limiting the mutational load.

---

\*aditi9783@gmail.com

# 1 Introduction

Genome sizes evolve by various mechanisms, some of which are common to all domains of life (insertions and deletions) while others are seen in some taxonomic groups more than others (horizontal gene transfer in bacteria and transposable element activity in eukaryotes). While one might think that genome expansion leads to the acquisition of more protein-coding genes and functions, genome size does not strongly correlate with organismal complexity (the C-value paradox). Whole-genome sequencing data provide some explanation for this paradox: appreciable variation in eukaryotic genome sizes has been attributed to ploidy [1], and to expansion of non-coding DNA such as introns, intergenic regions, and repeats [2]. Yet, genome size also positively correlates with the number of protein-coding genes [2], suggesting that larger genome size *is* a prerequisite for gaining new genes that could lead to phenotypic innovation.

Mutation rate, insertions and deletions (indels), and population size are three factors seen across the tree of life that are thought to influence genome size evolution. The negative correlation between genome size and point mutation rate is observed across the tree of life, from viruses to *Homo sapiens* [3]. However, a recent analysis based on more taxa found that this inverse relationship holds true only for prokaryotes and viruses, and that genome size and mutation rate are instead positively correlated in eukaryotes [4]. High point mutation rate forces viruses to maintain small genome sizes in an effort to limit the number of deleterious mutations [5]. This selection pressure to reduce genome size is so strong that viruses eliminate non-functional sequences inserted into their genomes [6] and lose an essential gene if it is transferred to the host genome [7]. This suggests that the point mutation rate and the evolution of genome size are inherently intertwined.

Population size, together with the point mutation rate and genome size, determines the mutation supply rate in an evolving population: if too many mutations are occurring, then reduction in any or all of point mutation rate, genome size, and population size can lower the mutation supply rate. Moreover, the effect of genetic drift is enhanced and purifying selection is weakened in small populations, allowing non-beneficial genome edits to persist for generations [8]. Lynch and Conery postulate that these—initially nonadaptive—edits can become a source of phenotypic innovation later on [2]. In symbiotic bacteria, small population size and asexual reproduction cause bacterial genomes to shrink to an extent that they are 2-4 times smaller than the smallest genome seen in an independent-living organism [9]. In contrast, large population sizes in microbial populations weaken the effect of random drift, preventing accumulation of non-functional DNA and genome growth [10].

In addition to point mutation rate and population size, biases in patterns of insertions and deletions (indel spectra) have been suspected to contribute to the variation in genome sizes we see today [11]. DNA loss via deletions is purported to be important in determining genome size, but this perspective is derived from analysis of a small number of eukaryotic genomes [12, 13]. Strong deletion bias was found in 12 bacterial species as well [14], the majority of which have

transposable element (TE) activity. Thus, it is likely that deletions outnumber insertions in taxa where TE proliferation leads to significant increases in non-functional DNA. This explanation, however, does not apply to genome size evolution in early living organisms and in taxa where TE activity is absent, and it is not clear how primordial genome editing mechanisms shaped the diversity in genome sizes we see today.

Digital evolution provides an apt platform for understanding the evolutionary processes that determine genome size. While naturally evolving biological systems can take a very long time to show observable changes, short generational time of digital organisms significantly reduces the time-scale of experiments to study evolutionary processes [15, 16, 17]. In the Avida artificial life platform, these digital organisms are simple computer programs that compete for resources to replicate via a mutation prone process (see Methods and Supplementary text), thus evolving under Darwinian dynamics [15, 16, 18]. The ability to control the mutation rate, genome sizes (length of the program), and population size allows inquiry into the impact of mutation rate and indel spectra on evolution of genome size. Avida has been previously used to test many evolutionary hypotheses that are difficult to test via biological experimental evolution, such as the evolution of genomic complexity [16], ‘survival of the flat-test’ effect in genotypes evolving at high mutation rates [19], co-evolution as a driving force for higher phenotypic complexity and evolvability [20], the time-dependent effect of genetic robustness on evolvability [21], and how standing genetic variation and environment influence evolutionary response to an environmental stimuli [22]. We used Avida to investigate genome size evolution because in addition to tracking genome edits and their fitness effects, it records evolution of phenotypic traits and thus can be used to interpret consequences of genome size evolution on phenotypic complexity.

Because avidians reproduce asexually and lack mechanisms of genome expansion such as TE activity, their evolutionary dynamics is most similar to that of viruses and prokaryotes. Thus, to examine the mechanisms of genome size evolution in asexual populations, we evolved populations of avidians at a range of mutation rates and followed the changes in their genome lengths, population fitness, genetic information, and phenotypic outcomes. Our results confirm that the genome size is negatively correlated with mutation rate. By tracking the changes in the genome size and the fitness effects of insertions and deletions that cause these changes, we find that insertions drive genome growth at low mutation rates, contributing to the evolution of phenotypic complexity via a two-step process: genome expansion followed by repurposing of the extra DNA to evolve new traits. Finally, we show that mutational load due to high mutation rate increases the selection pressure for reducing the genome size, resulting in smaller genomes with high information density. We conclude that genome size evolution is the result of a compromise between acquiring phenotypic complexity and restricting the mutational load.

## 2 Results and Discussion

### 2.1 Mutation rate is negatively correlated with genome size

We evolved avidians for 200,000 generations at six different point mutation rates. We found that genome sizes negatively correlate with the mutation rate (Fig. 1A; Spearman’s  $\rho = -0.72$ ,  $p < 3.6 \times 10^{-97}$ ). The mean population fitness also increased as the avidians’ genomes grew (Supplementary Fig. S1). The point mutation rates in our experiments ranged from  $2.5 \times 10^{-3}$  to 0.1, and the evolved genomic mutation rates ranged from 0.13 to 24.85 (genomic mutation rate was  $< 2$  for the lowest four point mutation rates). These genomic mutation rates are comparable to the ones seen in RNA viruses (0.025 in Influenza B virus, 1.1 in Hepatitis C Virus, and 4.6 in Bacteriophage Q $\beta$ ) [23]. Avidians did not evolve a constant genomic mutation rate in our experiments, as Drake observed in DNA microbes and RNA viruses [24, 25] and Knibbe et al reported in their digital evolution experiments [26]. A constant rate of genomic mutation is, however, not observed across the tree of life [3].

To test how genome size responds to changes in mutation rate, we switched the mutation rates of the avidians evolving at the lowest (0.0025) and the highest (0.1) point mutation rate after 100,000 generations. We find that the longer genomes that initially evolved at the low mutation rate began to shrink and those evolved at the high mutation rate began to expand (Fig. 1B), further establishing the direct influence of mutation rate on genomes size.

Since the ancestral genomes and population size were identical in all experiments, this negative correlation is independent of the effect of population size and the initial genomic content. By fixing the population size, we separated the influence of population size from that of mutation rate on genome size evolution, since it has been shown that population size influences genome size evolution as well [2].

### 2.2 Large genomes carry more genetic information

Although genome size does not correlate with organismal complexity (C-value paradox), complex organisms usually do have longer genomes. In other words, while genome expansion does not necessarily increase the number of functional sites in the genome, complex organisms are likely to have a higher amount of genetic information encoded into their genomes, which requires larger genomes. For example, even though *C. elegans* has a similar number of genes to *H. sapiens* (19,957 genes in the nematode compared to 20,181 in humans), the nematode has 20% less intergenic DNA and their mean intron size is 1/20th to that of humans [1]. On the premise that humans are more complex than *C. elegans*, one can argue that the expansion of non-coding DNA is at least partly responsible for this significant increase in complexity. Indeed,  $>85\%$  of the human genome is transcribed [27], contributing greatly to the non-coding RNA pool of the cell that regulates expression of protein-coding genes and participate in other

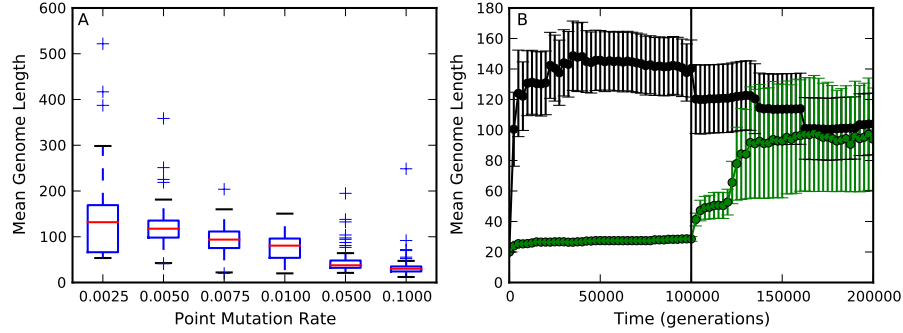


Figure 1: Point mutation rate is a strong determinant of genome size. A: Genome size and mutation rate are negatively correlated in asexual populations. The initial conditions, i.e. the ancestral genome and population size, were identical for all point mutation rates in our study (0.0025, 0.005, 0.0075, 0.01, 0.05, and 0.1). The avidian populations at the lowest mutation rate (0.0025) are still evolving (mean population fitness is still increasing, Supplementary Fig. S1) after 200,000 generations, explaining the higher variation in genome length for this mutation rate. Red lines are median values from 100 replicates, while the upper and lower bounds of the box are the third and first quartile, respectively. Whiskers are either 1.5 times the the quartile value or the extreme value in the data, whichever is closer to the median. Plus signs are outliers. B: The direct link between point mutation rate and genome size is further reinforced by switching the point mutation rate of population evolving at 0.0025 to 0.1 after 100,000 generations (black circles), and vice versa (green circles). The black line represents the generation where the mutation rates were switched. The long genomes shrink when mutation rate is increased and short genomes expand when mutation rate is decreased. Error bars represent  $\pm 1$  SE. Values represent the mean genome length across the population, averaged over 20 replicates.

cellular processes [28]. Even introns are not junk-DNA and contribute to the evolution of complexity in eukaryotes [29, 30]. About 20% of the pseudogenes are transcribed in humans [31], and are differentially expressed in cancers and viral infections [32, 33]. Thus, genome expansion, even if primarily of the non-coding DNA, likely increases the number of functional sites in the genome. Even if some of this inserted DNA is non-functional at the outset, evolution can repurpose it to achieve higher organismal complexity and genetic information [34, 35].

In our experiments, avidians that evolved long genomes at low mutation rates had higher genetic information (number of essential sites in the genome) than those that evolved at high mutation rates and had shorter genomes (Fig. 2; Spearman’s  $\rho = -0.86$ ,  $p < 6.4 \times 10^{-180}$ ). The longer genomes also evolved more traits (see Methods for an explanation of traits, and Supplementary Fig. S2), which are the computational equivalent of biological pathways that lead to observable phenotypes. The mean population fitness was also inversely related

to mutation rate, although the mean fitness of populations evolving at point mutation rate of 0.0025 was still increasing after 200,000 generations (Supplementary Fig. S1). This suggests that larger genome size is a necessary, if not sufficient, requirement for evolving phenotypic novelty. The avidians on average evolved fewer traits when the point mutation rate was switched half-way from 0.0025 to 0.1, and evolved more traits when mutation rate was switched from 0.1 to 0.0025, emphasizing the relationship between genome size, mutation rate, and phenotypic complexity (Fig. 1B and Supplementary Fig. S3).

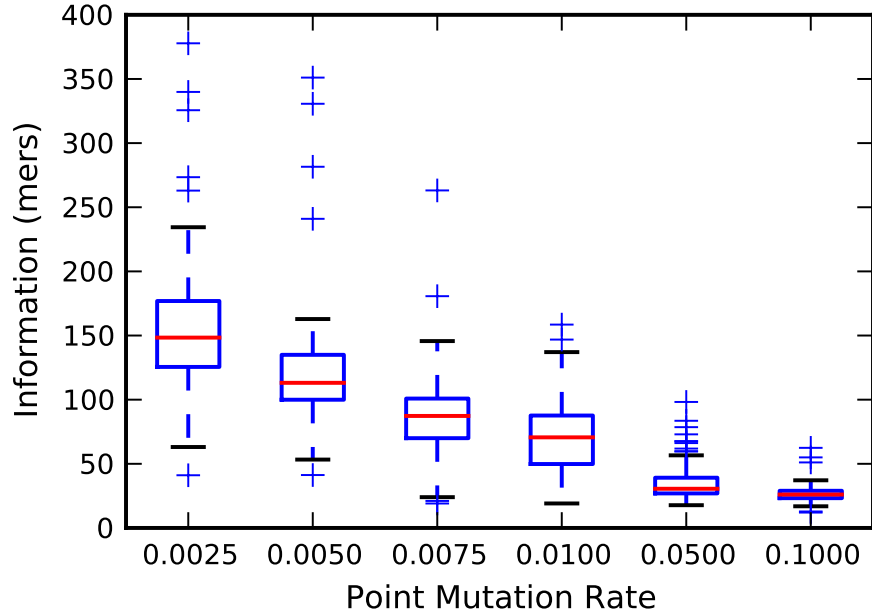


Figure 2: Genomes evolved at low mutation rates had higher genetic information (number of essential sites in the genome, see Methods) than genomes evolved at high mutation rates. The information measure is reported for the fittest genotype in each of the 100 replicate populations. Red lines are median values from 100 replicates, while the upper and lower bounds of the box are the third and first quartile, respectively. Whiskers are either 1.5 times the the quartile value or the extreme value in the data, whichever is closer to the median. Plus signs are outliers.

### 2.3 Beneficial insertions drive genome expansion at low mutation rates

To understand how genomes gain meaningful increases in size, we followed the genome edits (indels and mutations), the corresponding effect on fitness ( $s$ ), the

number of traits evolved, and genome size along the line of descent in *avidians* evolving at different mutation rates (Fig. 3). At the lowest point mutation rate in our experiments (Fig. 3A), the beneficial changes in the genome (green spikes) often align with evolution of new traits (blue line), as well as with insertions in the genome (red spikes). Insertions are largely beneficial compared to deletions at low mutation rate (Fig. 4). Phenotypic innovation (evolving a new trait) was preceded by insertion events 87% of the time (within the previous 20 ancestors along the line of descent), while deletions preceded innovation 60% of the time (null hypothesis: presence or absence of insertions is irrelevant to trait evolution, rejected with  $p < 1.0 \times 10^{-100}$ ,  $\chi^2$  test statistic =  $2.23 \times 10^5$ ). Thus, *avidian* genomes are likely to evolve new traits after an insertion event, suggesting that phenotypic innovation happens in a two-step process: genome expansion followed by evolution of a new trait by substitutions. Insertions are not deleterious per se (inset plots in Fig. 4) and thus persist in the line of descent. In fact, these inserted sequences may serve as substrates for evolving new phenotypic traits later on, contributing to increase in fitness and phenotypic complexity. In contrast, indels are infrequent at high mutation rates on the line of descent (Fig. 3B, also see Supplementary Fig. S4). As a result, the genomes do not grow and evolve fewer traits compared to the genomes evolved at low mutation rates.

This prominent role of beneficial insertions in genome evolution of asexual organisms is in contrast to how genome sizes are shaped by DNA loss in eukaryotes. The reported biases in indel spectra (rarity of long insertions and abundance of short deletions) are seen primarily in eukaryotic genomes [36]. Yet, a thermodynamic argument suggests that large indels are likely to increase genome size, since insertion events require only one breakpoint in the genome rendering large insertions less disruptive than large deletions [36, 13]. By the same argument, DNA loss is more likely to happen by small deletions to minimize the fitness cost to the organism. Thus, while eukaryotic genomes may evolve by rapid expansion due to whole genome duplication events and TE proliferation, asexual populations such as RNA viruses may have grown their genomes gradually via beneficial insertions. However, gradual increases in *avidian* genomes at low mutation rates is still followed by small deletions that fine-tune the genome size (Fig. 3).

## 2.4 High mutation rates force genomes to be small and informationally dense

If beneficial insertions drive genome expansion at low mutation rates, what keeps genomes small at high mutation rates? We find that the fitness cost of deleterious mutations is high at high mutation rates (Fig. 5A; Spearman's  $\rho = -0.71$ ,  $p < 1.1 \times 10^{-90}$ ). Since genotypes evolving at high mutation rates are compact, genetic information is forced to be distributed over a small number of sites (Fig. 5B), as in overlapping genes commonly seen in viral genomes. A deleterious mutation at a single such site can unfavorably affect multiple traits, increasing the overall fitness cost of deleterious mutations. Digital evolution

experiments also find that gene knockouts are more deleterious when pleiotropy is high, as is common in compact genomes [37]. Thus, not only is the mutational load high at high mutation rates, the deleterious mutations are costlier than they are at low mutation rates (mutational load is  $1 - e^{-\mu s}$ , as derived from [38]). This compounding factor only strengthens the selection pressure to decrease mutational load by reducing genome size, especially since population size is fixed in our experiments.

It should be noted that mutation rate can itself evolve to facilitate adaptation (reviewed in [3] and [39]). For example, the mutator strain of *E. coli* with a higher mutation rate than the wild-type bacteria showed the ability to adapt faster [40]. Even though the majority of mutations are deleterious, the ability to quickly find the adaptive beneficial mutations was enough to increase the population of the mutator strain relative to the wild-type [40]. However, this evolutionary advantage is short-lived and disappears once the beneficial mutations are found and there is no more fitness peak to climb [40, 41]. The mutator strain also does not propagate faster than the wild-type when a higher mutation supply is achieved by increasing the population size [40, 41]. Thus, environmental stresses such as starvation triggers a response in bacteria wherein mutation rate is elevated to quickly find beneficial mutations to adapt to the temporarily adverse conditions [42].

Since high mutation rate increases the mutational load in an evolving population, it makes sense that when the environmental stress is no longer present, the mutation rate would revert to the lower level. After all, the fitness cost of accumulating deleterious mutations would be too high if the rapid rate of adaptation afforded by high mutation rate is not needed. Mutator strains in well-adapted bacterial populations evolve decreased mutation rate as the opportunity for adaptation diminishes [43], an observation supported by digital evolution experiments [44]. Perhaps a continual need for adaptation is responsible for consistently high mutation rates in viruses, parasites, and sometimes in pathogenic bacteria where rapid adaptation to host immune responses is critical for surviving such an evolutionary arms race [45, 3, 46, 47]. The selection pressure to adapt quickly to a changing environment appears to trump the selection pressure to decrease mutational load by minimizing the mutation rate. However, mutational load can restrict virus adaptability due to an abundance of deleterious mutations [48]. Thus, the compromise between evolutionary forces for reducing the mutation load and maintaining high adaptability might shape the genome size and information density in RNA viruses.



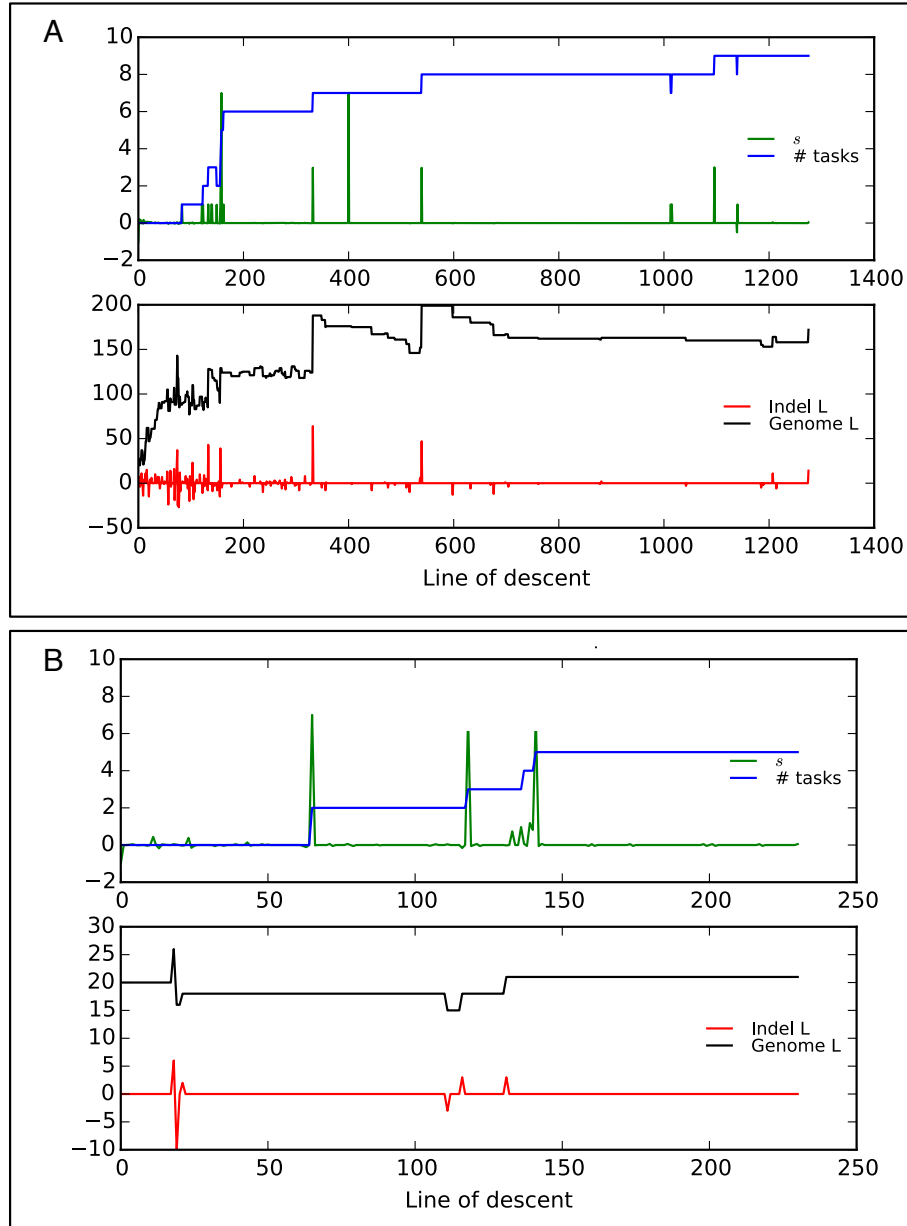


Figure 3: The line of descent (LOD) of the most fit genome is shown for a single replicate population evolving at the lowest (0.0025, A) and the highest (0.1, B) point mutation rate in our study. The fitness effects of genome edit events (insertions, deletions, base substitutions) are shown in green, the number of evolved traits is shown in blue, the size of indels is shown in red, and the genome length is shown in black. At low mutation rate (top panel, A), new traits (in blue) often evolved following beneficial genomic events (green spikes), and are sometimes concurrent with insertion events (red spikes). These beneficial insertions appear to increase the genome size (black line) over time. At the high mutation rate (bottom panel, B), insertion events are not as frequent as at low mutation rates (also see Supplementary Fig. S4), with genome size staying relatively constant. The line of descent (LOD) maps for other mutation rates can be found in Supplementary Fig. S5.

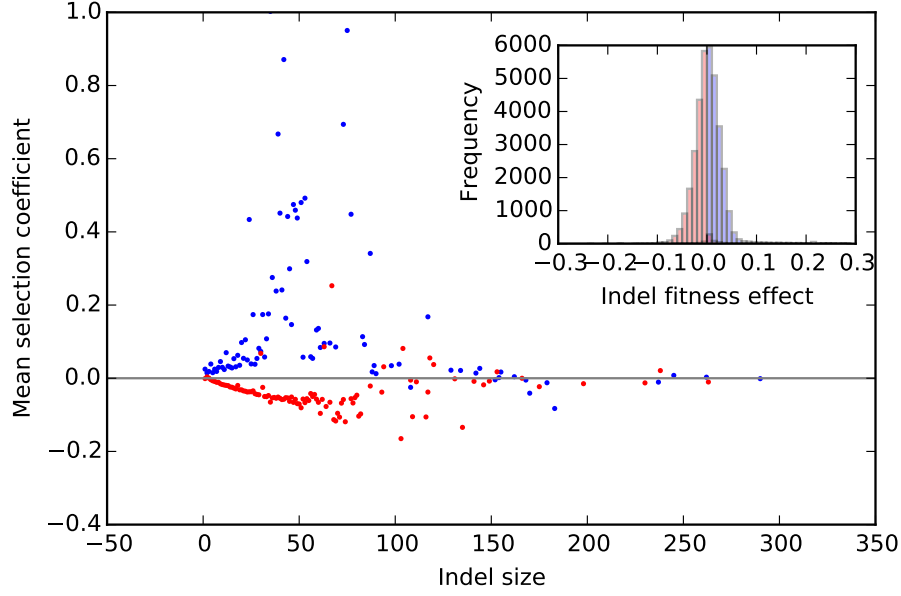


Figure 4: The average fitness effect of insertions (blue) and deletions (red) as a function of indel size is shown for 100 replicate populations evolving at the point mutation rate of 0.0025. Indels above the gray line ( $s = 0$ ) are beneficial and those below the gray line are deleterious. Small insertions (blue dots) are usually beneficial, while small deletions (red dots) are usually deleterious. The inset plot shows the histograms of fitness effects of insertions (blue bars, total 19,262 insertions) and deletions (red bars, total 16,998 deletions) along the line of descent in 100 replicate populations. Insertions (blue bars) are usually beneficial (*i.e.*, fitness effect  $> 0$ ), and deletions (red bars) are usually deleterious (fitness effect  $< 0$ ). The two distributions are significantly different (Kolmogorov-Smirnov two-sided test,  $p < 1 \times 10^{-100}$ ).

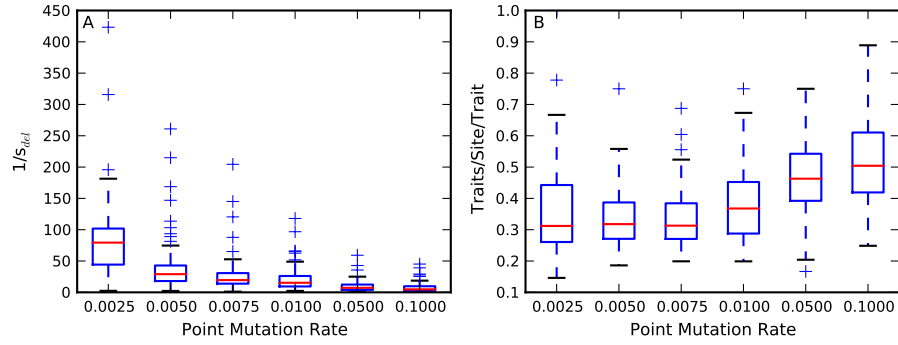


Figure 5: Deleterious mutations at high mutations rates are more costly due to informationally dense genomes. The inverse of the harmonic mean of deleterious selection coefficients for the fittest genotype from each replicate shows that deleterious mutations are costlier at high mutation rates (A). This can be explained by the high coding density in these genomes (B). Traits/site/trait represents how many traits are encoded per site, normalized by the total number of evolved traits, and thus is a measure of coding density of the genome. Red lines are median values from 100 replicates, while the upper and lower bounds of the box are the third and first quartile, respectively. Whiskers are either 1.5 times the the quartile value or the extreme value in the data, whichever is closer to the median. Plus signs are outliers.

### 3 Conclusions

While empirical studies reveal significant aspects of genome size evolution, digital evolution systems provide an opportunity to observe evolution-in-action and to manipulate evolutionary parameters in ways that allows exploring the relative importance of the many evolutionary forces that simultaneously act on genomes. Comparative genomics analyses have unearthed important relationships between population size, mutation rate, gene content, genome size, and their combined influence on evolution of complexity. However, digital evolution experiments complement these retrospective observations by investigating evolutionary processes that are difficult to test experimentally. In our experiments at a range of mutation rates, we find concurrence with the empirical finding that the point mutation rate is negatively correlated with genome size. By tracking the genomes along the line of descent, we find insertions to be significantly beneficial compared to deletions, suggesting that before the advent of complex mechanisms of genome edits such as TE activity, beneficial insertions drove genome expansion. That these insertions are followed by phenotypic innovations further explains why insertions are evolutionarily favored in asexual populations. At the same time, the point mutation rate influences genome size via the mutational load. Thus, unless high mutation rate provides a critical evolutionary advantage such as rapid adaption to a temporary environmental stress, the selection pressure to reduce mutational load forces the genomes to shrink at high mutation rates. This genome shrinkage results in genomes packed with genetic information, and this compactness likely increases the fitness cost of deleterious mutations, further compounding the severity of mutational load. Still, a high point mutation rate is frequently seen in natural populations, especially in viruses, suggesting that the selection pressure to maintain high evolvability (for example, against a highly adaptive host immune system) can take precedence over selection pressure to reduce mutational load in the fight to survival.

The evolution of genome size is a complex phenomenon, especially in eukaryotes due to TE activity and expansion of non-coding DNA. Our analyses of asexual populations evolving at fixed point mutation and indel rates reveal the fundamental roles that indel spectra and mutational load play in determining genome size and phenotypic diversity. Investigations into eukaryotic genome size evolution by including recombination and TE activity in digital evolution platforms will allow comparisons with asexual genome size evolution, and can shed light on evolution of complex genome editing mechanisms.

## 4 Materials and Methods

### 4.1 Avida digital evolution platform

Avida is a digital evolution platform which provides an environment within which digital organisms, using sets of instructions analogous to codons, experience selective pressures to develop genes that encode logical operations [49, 50,

15]. Performing these operations provides these avidians with single instruction processing units or SIPs, their energy currency equivalent of ATP. By performing increasingly complex boolean logic calculations, the avidians are able to accrue larger amounts of energy to outcompete their neighbors, much as living organisms participate in the evolutionary arms race in their own ecological niches. They replicate by error-prone mechanisms, thus mimicking Darwinian evolution. Since we investigated qualities that are innate to genomes as stores of information, and are not mechanism dependent (other than a requirement for a lack of total fidelity in replication), Avida is an ideal model system to study evolutionary forces that drive genome evolution in asexual populations.

## 4.2 Experimental Design

To test the role of the mutation rate in driving genome size evolution, we evolved 100 replicate populations at various point mutation rates ( $\mu = \{0.0025, 0.005, 0.0075, 0.01, 0.05, 0.1\}$ ) for 200,000 generations. Insertions and deletions occurred with equal frequency at a constant rate of 0.05 indels per generation. Indel size was uniformly distributed, with genome size changing at most by 10% in any given generation. All populations were initialized with an identical ancestral genome of size 20. Population size was fixed at the default 3600 individuals. There was no structure in the evolving populations (i.e. a well-mixed environment). An additional 40 populations were evolved for 200,000 generations where the mutation rates were switched after 100,000 generations as follows: 20 populations that initially evolved at a point mutation rate of 0.0025 were switched to a point mutation rate of 0.1 after 100,000 generations, and the remaining 20 populations were switched from point mutation rate of 0.1 to 0.0025 after 100,000 generations.

## 4.3 Line of Descent

To track the effect of genome edits on genome size and phenotypic evolution, we analyzed the Line of Descent (LOD) of the fittest individual from each replicate population at the end of the evolution experiments. A LOD is a lineage of every ancestor of the evolved genotype that had the highest fitness at the end of 200,000 generations. It tracks every genome edit (and its corresponding effect on fitness) that was fixed in the lineage. This genotypic “fossil record” allows identifying those mutations that lead to evolutionary innovations and determine the respective role of insertions and deletions in genome size evolution.

## 4.4 Data Analysis

We calculated statistics at both the population level and for individual genotypes. The mean genome length and the mean fitness was calculated by averaging the relevant values across all genotypes in each population which was then averaged over 100 replicate populations. For the rest of our reported data, we calculated statistics from the fittest genotype in the final evolved population. A

genotype’s information content was estimated as  $I = L - \sum_i^L \log_{26} \nu(i)$ , where  $L$  is the genome size, 26 is the alphabet size for avidian genomes, and  $\nu(i)$  is the number of mutations that are neutral or beneficial (see [16] for further explanation of this estimation). Thus, information content is a measure of the number of essential sites in a genome. The number of phenotypic traits a genotype possesses is calculated as the number of different boolean logic calculations it can perform.

The traits per site per trait measure is determined by performing knockout mutations at every site in the genome and then counting the number of traits that are lost due to each knockout mutation (lethal knockouts are not considered). This gives the number of traits that utilize each genomic site, and average of this quantity over the length of the genome gives the overall number of traits encoded per site. The normalized trait/site/trait is then calculated by dividing the traits/site by the total number of traits evolved by the genome.

## 5 Acknowledgments

This work was supported in part by the National Science Foundation’s BEACON Center for the Study of Evolution in Action, under contract No. DBI-0939454. We wish to acknowledge the Michigan State University High Performance Computing Center and the Institute for Cyber Enabled Research for computational support. Michael Miyagi thanks the Freshman Research Initiative program at the University of Texas at Austin for undergraduate research opportunity.

## References

- [1] Ryan J Taft, Michael Pheasant, and John S Mattick. The relationship between non-protein-coding dna and eukaryotic complexity. *Bioessays*, 29(3):288–99, Mar 2007.
- [2] Michael Lynch and John S Conery. The origins of genome complexity. *Science*, 302(5649):1401–4, Nov 2003.
- [3] P D Sniegowski, P J Gerrish, T Johnson, and A Shaver. The evolution of mutation rates: separating causes from consequences. *Bioessays*, 22(12):1057–66, Dec 2000.
- [4] Michael Lynch. Evolution of the mutation rate. *Trends Genet*, 26(8):345–52, Aug 2010.
- [5] Edward C Holmes. Error thresholds and the constraints to rna virus evolution. *Trends Microbiol*, 11(12):543–6, Dec 2003.
- [6] Mark P Zwart, Anouk Willemsen, José-Antonio Daròs, and Santiago F Elena. Experimental evolution of pseudogenization and gene loss in a plant rna virus. *Mol Biol Evol*, 31(1):121–34, Jan 2014.

- [7] Nicolas Tromas, Mark P Zwart, Javier Forment, and Santiago F Elena. Shrinkage of genome size in a plant rna virus upon transfer of an essential viral gene into the host genome. *Genome Biol Evol*, 6(3):538–50, Mar 2014.
- [8] Chih-Horng Kuo, Nancy A Moran, and Howard Ochman. The consequences of genetic drift for bacterial genome complexity. *Genome Res*, 19(8):1450–4, Aug 2009.
- [9] John P McCutcheon and Nancy A Moran. Extreme genome reduction in symbiotic bacteria. *Nat Rev Microbiol*, 10(1):13–26, Jan 2012.
- [10] Michael Lynch. Streamlining and simplification of microbial genome architecture. *Annu Rev Microbiol*, 60:327–49, 2006.
- [11] Alexander E Vinogradov. Evolution of genome size: multilevel selection, mutation bias or dynamical chaos? *Curr Opin Genet Dev*, 14(6):620–6, Dec 2004.
- [12] D A Petrov, T A Sangster, J S Johnston, D L Hartl, and K L Shaw. Evidence for dna loss as a determinant of genome size. *Science*, 287(5455):1060–2, Feb 2000.
- [13] T Ryan Gregory. Insertion-deletion biases and the evolution of genome size. *Gene*, 324:15–34, Jan 2004.
- [14] Chih-Horng Kuo and Howard Ochman. Deletional bias across the three domains of life. *Genome Biol Evol*, 1:145–52, 2009.
- [15] Charles Ofria and Claus O Wilke. Avida: a software platform for research in computational evolutionary biology. *Artif Life*, 10(2):191–229, 2004.
- [16] C Adami, C Ofria, and T C Collier. Evolution of biological complexity. *Proc Natl Acad Sci U S A*, 97(9):4463–8, Apr 2000.
- [17] Bérénice Batut, David P Parsons, Stephan Fischer, Guillaume Beslon, and Carole Knibbe. In silico experimental evolution: a tool to test evolutionary scenarios. *BMC Bioinformatics*, 14 Suppl 15:S11, 2013.
- [18] Richard E Lenski, Charles Ofria, Robert T Pennock, and Christoph Adami. The evolutionary origin of complex features. *Nature*, 423(6936):139–44, May 2003.
- [19] C O Wilke, J L Wang, C Ofria, R E Lenski, and C Adami. Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature*, 412(6844):331–3, Jul 2001.
- [20] Luis Zaman, Justin R Meyer, Suhas Devangam, David M Bryson, Richard E Lenski, and Charles Ofria. Coevolution drives the emergence of complex traits and promotes evolvability. *PLoS Biol*, 12(12):e1002023, Dec 2014.

- [21] Santiago F Elena and Rafael Sanjuán. The effect of genetic robustness on evolvability in digital organisms. *BMC Evol Biol*, 8:284, 2008.
- [22] Daniel R O’Donnell, Abhijna Parigi, Jordan A Fish, Ian Dworkin, and Aaron P Wagner. The roles of standing genetic variation and evolutionary history in determining the evolvability of anti-predator strategies. *PLoS One*, 9(6):e100163, 2014.
- [23] Rafael Sanjuán, Miguel R Nebot, Nicola Chirico, Louis M Mansky, and Robert Belshaw. Viral mutation rates. *J Virol*, 84(19):9733–48, Oct 2010.
- [24] J W Drake and J J Holland. Mutation rates among rna viruses. *Proc Natl Acad Sci U S A*, 96(24):13910–3, Nov 1999.
- [25] J W Drake. A constant rate of spontaneous mutation in dna-based microbes. *Proc Natl Acad Sci U S A*, 88(16):7160–4, Aug 1991.
- [26] Carole Knibbe, Guillaume Beslon, Virginie Lefort, F Chaudier, and J-M Fayard. Self-adaptation of genome size in artificial organisms. In *Advances in Artificial Life*, pages 423–432. Springer, 2005.
- [27] Matthew J Hangauer, Ian W Vaughn, and Michael T McManus. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding rnas. *PLoS Genet*, 9(6):e1003569, Jun 2013.
- [28] Igor Ulitsky and David P Bartel. lincrnas: genomics, evolution, and mechanisms. *Cell*, 154(1):26–46, Jul 2013.
- [29] J S Mattick and M J Gagen. The evolution of controlled multitasked gene networks: the role of introns and other noncoding rnas in the development of complex organisms. *Mol Biol Evol*, 18(9):1611–30, Sep 2001.
- [30] Igor B Rogozin, Liran Carmel, Miklos Csuros, and Eugene V Koonin. Origin and evolution of spliceosomal introns. *Biol Direct*, 7:11, 2012.
- [31] Deyou Zheng, Adam Frankish, Robert Baertsch, Philipp Kapranov, Alexandre Reymond, Siew Woh Choo, Yontao Lu, France Denoeud, Stylianos E Antonarakis, Michael Snyder, Yijun Ruan, Chia-Lin Wei, Thomas R Gingeras, Roderic Guigó, Jennifer Harrow, and Mark B Gerstein. Pseudogenes in the encode regions: consensus annotation, analysis of transcription, and evolution. *Genome Res*, 17(6):839–51, Jun 2007.
- [32] Laura Polisenio, Leonardo Salmena, Jiangwen Zhang, Brett Carver, William J Haveman, and Pier Paolo Pandolfi. A coding-independent function of gene and pseudogene mrnas regulates tumour biology. *Nature*, 465(7301):1033–8, Jun 2010.
- [33] Aditi Gupta, C. Titus Brown, Yong-Hui Zheng, and Christoph Adami. Differentially-expressed pseudogenes in hiv-1 infection. *Viruses*, 7(10):5191–5205, 2015.



- [34] Michael Lynch. The frailty of adaptive hypotheses for the origins of organismal complexity. *Proc Natl Acad Sci U S A*, 104 Suppl 1:8597–604, May 2007.
- [35] Carole Knibbe, Antoine Coulon, Olivier Mazet, Jean-Michel Fayard, and Guillaume Beslon. A long-term evolutionary pressure on the amount of noncoding dna. *Mol Biol Evol*, 24(10):2344–53, Oct 2007.
- [36] Dmitri A Petrov. Mutational equilibrium model of genome size evolution. *Theor Popul Biol*, 61(4):531–44, Jun 2002.
- [37] Carole Knibbe, Olivier Mazet, Fabien Chaudier, Jean-Michel Fayard, and Guillaume Beslon. Evolutionary coupling between the deleteriousness of gene mutations and the amount of non-coding sequences. *J Theor Biol*, 244(4):621–30, Feb 2007.
- [38] M Kimura and T Maruyama. The mutational load with epistatic gene interactions in fitness. *Genetics*, 54(6):1337–51, Dec 1966.
- [39] D Metzgar and C Wills. Evidence for the adaptive evolution of mutation rates. *Cell*, 101(6):581–4, Jun 2000.
- [40] A Giraud, I Matic, O Tenaillon, A Clara, M Radman, M Fons, and F Taddei. Costs and benefits of high mutation rates: adaptive evolution of bacteria in the mouse gut. *Science*, 291(5513):2606–8, Mar 2001.
- [41] J A Arjan, M Visser, C W Zeyl, P J Gerrish, J L Blanchard, and R E Lenski. Diminishing returns from mutation supply rate in asexual populations. *Science*, 283(5400):404–6, Jan 1999.
- [42] S M Rosenberg, C Thulin, and R S Harris. Transient and heritable mutators in adaptive evolution in the lab and in nature. *Genetics*, 148(4):1559–66, Apr 1998.
- [43] Sébastien Wielgoss, Jeffrey E Barrick, Olivier Tenaillon, Michael J Wiser, W James Dittmar, Stéphane Cruveiller, Béatrice Chane-Woon-Ming, Claudine Médigue, Richard E Lenski, and Dominique Schneider. Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load. *Proc Natl Acad Sci U S A*, 110(1):222–7, Jan 2013.
- [44] Jeff Clune, Dusan Misevic, Charles Ofria, Richard E Lenski, Santiago F Elena, and Rafael Sanjuán. Natural selection fails to optimize mutation rates for long-term adaptation on rugged fitness landscapes. *PLoS Comput Biol*, 4(9):e1000187, 2008.
- [45] Philip B Greenspoon and Leithen K M’Gonigle. The evolution of mutation rate in an antagonistic coevolutionary model with maternal transmission of parasites. *Proc Biol Sci*, 280(1761):20130647, Jun 2013.

- [46] Christina Hoboth, Reinhard Hoffmann, Anja Eichner, Christine Henke, Sabine Schmoldt, Axel Imhof, Jürgen Heesemann, and Michael Hogardt. Dynamics of adaptive microevolution of hypermutable *pseudomonas aeruginosa* during chronic pulmonary infection in patients with cystic fibrosis. *J Infect Dis*, 200(1):118–30, Jul 2009.
- [47] Santiago F Elena and Rafael Sanjuán. Adaptive value of high mutation rates of rna viruses: separating causes from consequences. *J Virol*, 79(18):11555–8, Sep 2005.
- [48] Oliver G Pybus, Andrew Rambaut, Robert Belshaw, Robert P Freckleton, Alexei J Drummond, and Edward C Holmes. Phylogenetic evidence for deleterious mutation load in rna viruses and its contribution to viral evolution. *Mol Biol Evol*, 24(3):845–52, Mar 2007.
- [49] Chris Adami and C. Titus Brown. Evolutionary learning in the 2d artificial life system ‘avida’. *Artificial life IV.*, 1194:377–381, 1994.
- [50] Christoph. Adami. *Introduction to artificial life*, volume 1. Springer Science & Business Media, 1998.

# Supplementary Materials

## Expanded Methods: Avida

Avida is a digital experimental evolution platform where populations of simple computer programs (avidians) compete for the resources needed to self-replicate via error-prone mechanisms. The avidian genome consists of computer instructions which are executed during its life cycle to perform boolean logic calculations as well as to replicate its genome. Since evolution in Avida comprises genetic variation affecting ability to evolve phenotypic traits and to replicate, differential fitness dependent on this heritable variation and competition for computational resources causes avidians to undergo natural selection comparable to biological populations.

The Avida world consists of a 60x60 toroidal grid with at most one avidian per cell, resulting in a fixed population size of 3600. Each child avidian is placed in any one of the 3600 cells after successful replication (although new offspring are preferentially placed in empty cells if available), making the population well-mixed. When the population is at its carrying capacity, the avidian occupying the cell chosen for a new offspring will be removed from the population. This random selection of individuals for removal adds an element of genetic drift to avidian populations.

Absolute time in Avida is divided into updates. During each update, the population executes 30N instructions, where N is the population size. The ability to execute these instructions (comparable to energy units in cells—ATP), called Single Instruction Processing Units (SIPs), are distributed across the population. How these SIPs get distributed among the individuals in the population is dependent on a characteristic possessed by each individual called merit. In a monoclonal population, every individual will possess on average 30 SIPs per update. However, if one individual has a greater merit than others in the population, it is expected to receive more SIPs per update than the other individuals. This allows it to execute and copy its genome faster than other individuals. Therefore, as reproduction speed is the primary target of selection in this type of simple environment, increased merit results in increased fitness, and organisms with an increased merit will be under positive selection. In our experiments, we record data every generation, starting from the ancestral population, which marks generation 0. All progeny of the ancestral population constitute generation 1, and so forth.

Avidians increase their merit through the evolution of phenotypic traits. These traits are the ability to perform boolean logic computations. In the default Avida environment, the Logic-9 environment [18], populations can evolve up to 9 of these traits. Performing these traits result in a multiplicative increase in an individual's merit (ranging from a multiple of 2 for simple traits to 32 for the most complex trait). The evolution of these traits require many point mutations and a genome size large enough to contain the instructions necessary to perform these computations. Because these traits increase merit, and thus replication speed, the evolution of these traits are also under strong selection.

Each individual can perform each trait once during their lifespan, and there is no limit to the number of times a trait can be performed in a population. Because an individual's performance of a trait does not limit the others in the population, there is only one niche in the environment. Therefore, fitness is frequency-independent.

During an avidian's lifespan, it will eventually start to undergo genome replication. As it copies its genome's instructions into a blank daughter genome, some instructions may be copied inaccurately at a point mutation rate set by the experimenter. Additionally, insertion and deletion mutations can occur either during genome replication or during genome division into new daughter genomes. In the experiments performed here, insertion and deletion mutations (indels) were enacted upon genome division. Genome sizes can change every generation by at most 10% (the default is a maximum change of 100%). For every indel, two spots in the genome were randomly selected. If the indel was a deletion, everything between those two spots was deleted. If the indel was an insertion, that section of the genome was duplicated. Insertions and deletions occurred at equal frequencies in our experiments.

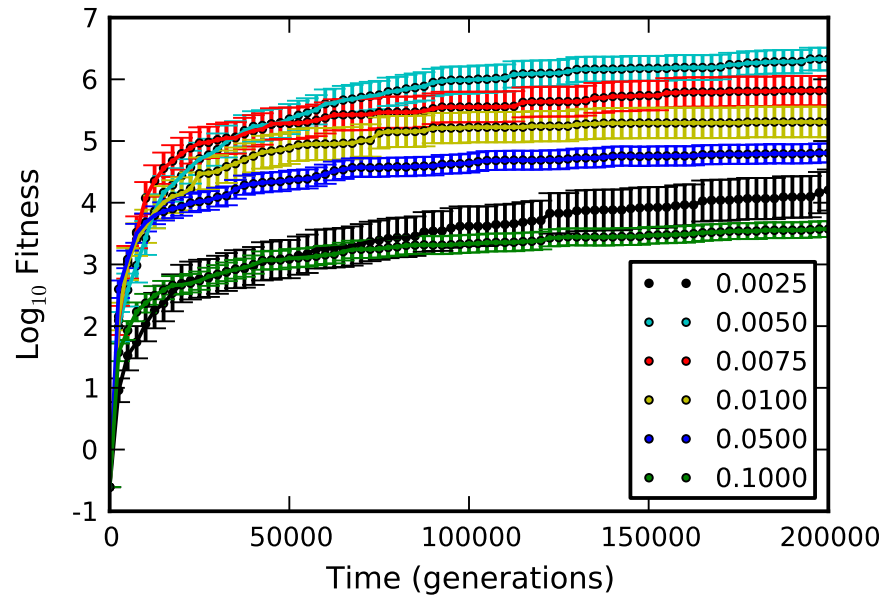


Figure S1: Increase in population fitness over 200,000 generations is shown for six point mutation rates (0.0025, 0.005, 0.0075, 0.01, 0.05, and 0.1). The  $\log_{10}$  of fitness is averaged over 100 replicate populations. Error bars represent  $\pm 1$  SE.

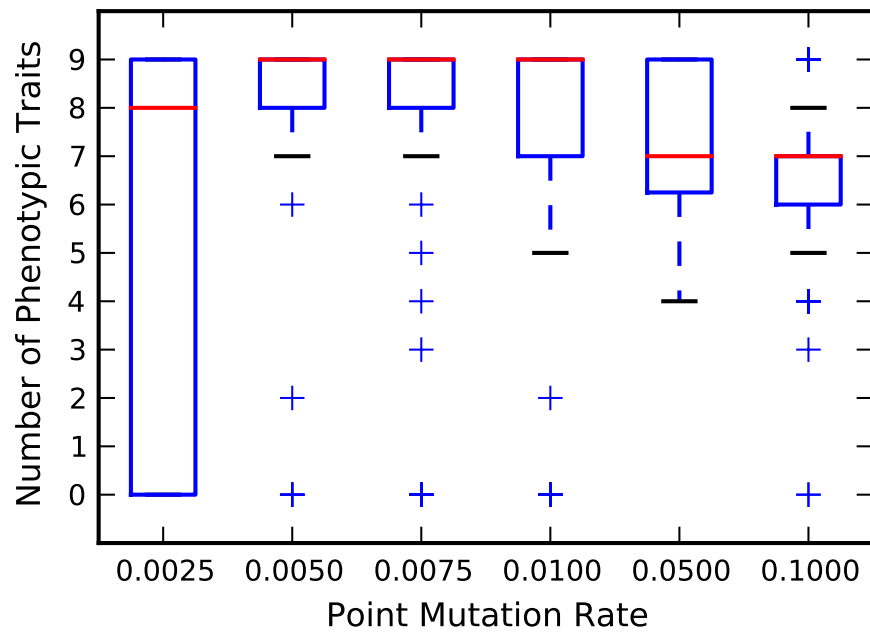


Figure S2: Number of traits evolved by avidians at six different point mutation rates (maximum number of traits that can be evolved is 9). Red lines are median values from 100 replicate populations, while the upper and lower bounds of the box are the third and first quartile, respectively. Whiskers are either 1.5 times the the quartile value or the extreme value in the data, whichever is closer to the median. Plus signs are outliers.

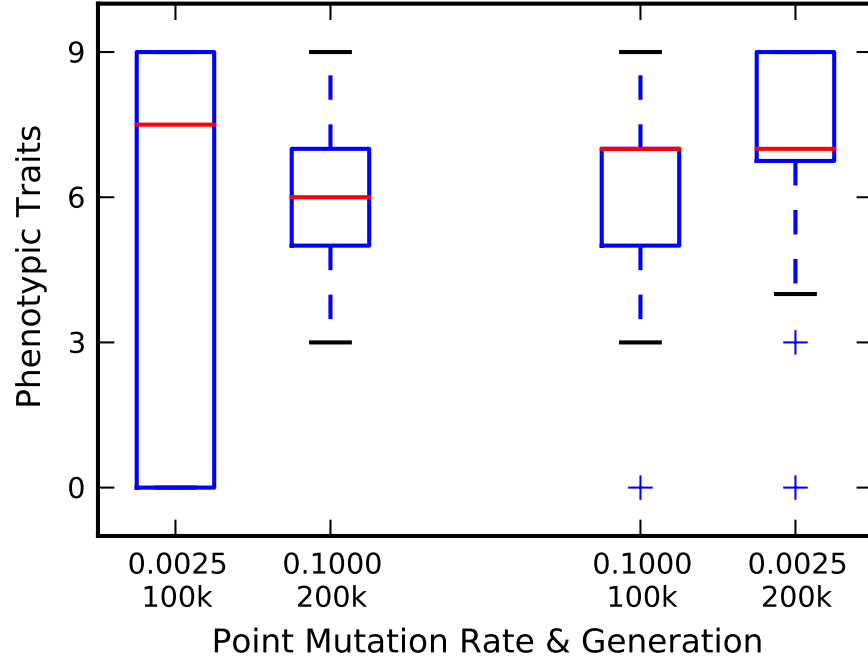


Figure S3: The number of traits evolved by avidians after mutation rate was switched at the mid-point (at 100,000 generations) in the 200,000 generation long simulations. Left side of the figure shows the statistics for number of traits evolved by the population evolving at point mutation rate of 0.0025, and the number of traits the same population evolved after evolving at mutation rate of 0.1 for 100k generations. Right side shows the reverse scenario: traits evolved by population evolving at point mutation rate of 0.1 at 100k generations, and after mutation rate is switched to 0.0025 for additional 100k generations. Red lines are median values from 20 replicate populations, while the upper and lower bounds of the box are the third and first quartile, respectively. Whiskers are either 1.5 times the the quartile value or the extreme value in the data, whichever is closer to the median. Plus signs are outliers.

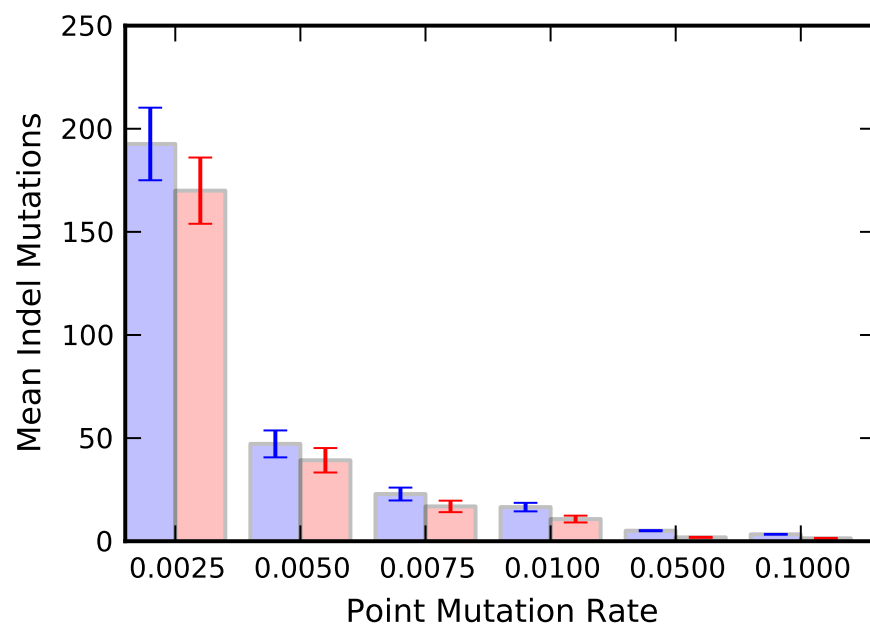


Figure S4: Insertions (blue bars) are more common than deletions (red bars) at all mutation rates, and frequency of indels decreases as mutation rate increases. Average values over 100 replicates is reported with error bars showing  $\pm 1$  SE.



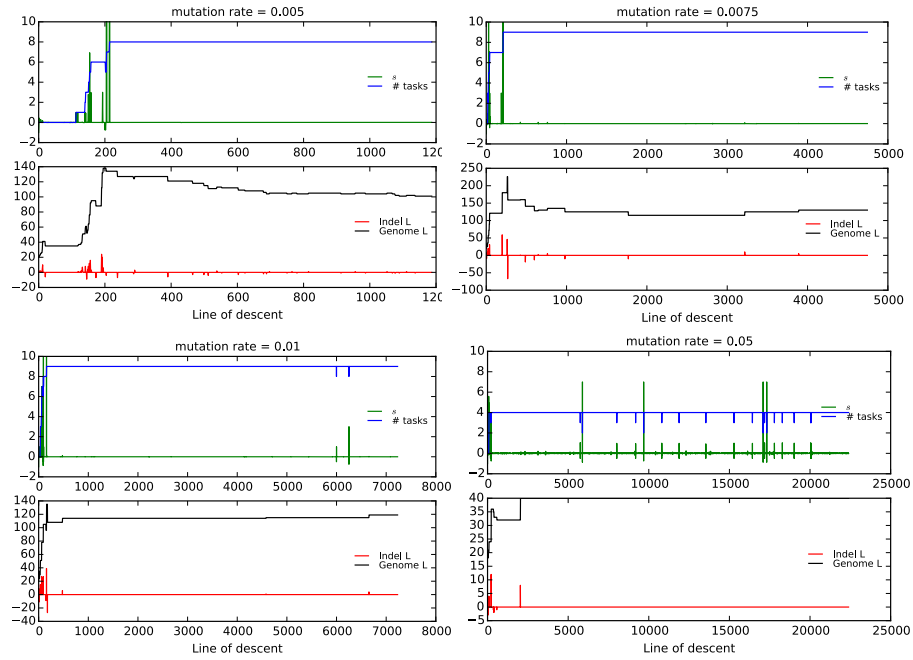


Figure S5: The line of descent (LOD) of the most fit genome is shown for a single replicate population evolving at the point mutation rates 0.005, 0.0075, 0.01, and 0.05. The fitness effects of genome edit events (insertions, deletions, base substitution) is shown in green, the number of traits evolved over time is shown in blue, the size of indels is shown in red, and the genome length is shown in black.